ALGO BASICO SOBRE LOS INSTRUMENTOS DE MEDIDA: VALIDEZ, FIABILIDAD, SENSIBILIDAD Y ESPECIFICIDAD.

SOMETHING BASIC ABOUT MEASURING INSTRUMENTS: VALIDITY, ACCURACY, SENSITIVENESS AND SPECIFICITY.

Manuel Castro Bouzas. Psicólogo Clínico Servizo de Psiquiatría Complejo Hospitalario Arquitecto Marcide - Profesor Nóvoa Santos (Ferrol), manuel.castro.bouzas@sergas.es

Abstract: In this article basic topics about the main properties of measuring instruments are provided. In accordance with this objective, concepts on validity, accuracy, sensitiveness and specificity are explained and likewise, their relations with concrete examples on them are given. After defining these concepts, namely from a practical point of view, it is remarked the need of using only instruments fulfilling those required properties.

Resumen: En este artículo se dan los conceptos básicos sobre las principales propiedades de los instrumentos de medida. Siguiendo ese objetivo se explican los conceptos de validez, fiabilidad, sensibilidad y especificidad y se proporcionan ejemplos concretos sobre los mismos. Tras definir los conceptos, principalmente desde un punto de vista práctico, se resalta la

necesidad de usar unicamente instrumentos que cumplan adecuadamente con estos requisitos.

Terms: Accuracy, specificty, validity, sensitivity, measuring instruments.

Términos: Validez, fiabilidad, sensibilidad, especificidad, instrumentos de medida.

A modo de preliminar

Cuando se me planteo la posibilidad de escribir un artículo para una revista de terapia ocupacional la verdad es que por mi cabeza pasaron varias preguntas. La más perentoria tenía que ver con el tema y el interés del mismo para los lectores. Y pasaron los días, y las llamadas insistentes del compañero que había contactado conmigo no hacían más que acrecentar el arrepentimiento por haber dicho que sí a aquella generosa oferta. Las musas venían, y los borradores se sucedían.

Al final me quedé con uno de los temas sobre los que más he dialogado con los estudiantes de Terapia Ocupacional (y también de otras disciplinas) cuando rotaban por el Servicio de Psiquiatría: ¿qué es eso de los tests? ¿qué es eso de validez y otras cosas que oigo mencionar sin saber muy bien qué significan?

Instrumentos de medida

Yo siempre “atacaba” de la misma manera: ¿qué es para vosotros un test? Esta pregunta, que tan pocos triunfos me ha deparado a la hora de ligar, permitía la expresión de esa asociación de índole social entre las palabras

psicología y test, tan frecuente como incompleta. Pero de este modo podía introducir el concepto que realmente me interesaba: los instrumentos de medida. Ante estos, tan frecuentes en la bibliografía y manuales de texto, la respuesta más frecuente era encogerse los hombros y dar una respuesta genérica del tipo "¡Ah, esos!", que muy bien no sabía cómo interpretar. De todos modos si no desean leer este artículo tendrían esta información en otras obras como Cronbach (1), Muñoz Rodríguez (2) y Cone (3)

Así que tenía que empezar por definir que se podía entender por instrumento de medida. Quizás una definición útil de instrumento de medida sea la de un conjunto de elementos que permiten asignar un valor numérico a una determinada dimensión de un objeto o ser. Ni más, pero tampoco menos.

Retirado del diálogo la palabra test, por el significado social no muy bien aclarado que tiene y que podría incitar más a la confusión que a la claridad, e introducido la expresión "Instrumento de medida" sigamos introduciendo más elementos necesarios.

¿Cualquier conjunto de elementos puede ser un instrumento de medida? En mi opinión no, y esta negativa es en el campo de las ciencias humanas más taxativa. En las ciencias físicas los instrumentos de medida deben reunir la pertinencia inequívoca entre sus características y lo medido. De ese modo a nadie, salvo despistadas excepciones, se le podría ocurrir utilizar una botella con la medida de un litro para medir una distancia. Y dejarlo así (alguno habrá pensado en hacerlo, y posteriormente medir la botella con una regla y

hacer la transformación a una unidad de medida de longitud como podrían ser los centímetros... pero eso sería complicarse la vida y no afecta al elemento fundamental de la pertinencia. Así que dejaos de experimentos y comprad un metro). De hacerlo sería una aplicación salvaje en este campo de uno de los Principios de Peter (Cuando tenemos que clavar algo todo aquello que tengamos en la mano cumplirá a función de martillo).

Esto, tan claro en las ciencias físicas, se obvia en las humanas. De este modo parece que cualquier conjunto de elementos sería susceptible de convertirse en un instrumento de medida. ¿Estáis seguros? ¿Podemos estar seguros de decir algo con sentido real cuando nosotros afirmamos que un resultado numérico representa, por ejemplo, la "orientación" de una persona en su entorno o su "independencia en las actividades de la vida diaria"? Aquí es cuando se introducen los conceptos de validez y fiabilidad.

Validez y fiabilidad

Antes, cuando mencionaba el tema de la botella y la distancia utilizaba la palabra pertinencia. La verdad es que fue deliberado para evitar el uso de la palabra validez. La misma es la palabra adecuada para indicar aquella propiedad de los instrumentos de medida que hace referencia a la relación entre un resultado numérico y un criterio que tomamos como patrón oro, con el que se establecerá una comparación. Este patrón oro es fundamental, ya que será el criterio de "realidad" que se supone el resultado numérico obtenido debe representar. De este modo en las fichas técnicas de los

instrumentos debe constar con qué criterio se comparó la puntuación obtenida y la correlación obtenida (se asume que ésta debe ser alta, y que cuanto más alta, más válido es el test).

Lo anterior es una simplificación. Con frecuencia los instrumentos surgen de una teoría o modelo que por sí definen algunas características que obligatoriamente hay que considerar, y no otras. Y eso nos puede llevar a la adecuación de los patrones oro, etc... Y de igual manera nosotros podemos comparar la puntuación con un criterio en el mismo momento (validez concurrente) o con otro que se medirá más adelante (predictiva). Pero no pretendo liar más el tema. Por ejemplo, tomemos que estamos interesados una escala que vaya a medir algo que desde nuestro modelo se puede denominar "autonomía personal". En esta dimensión nuestro modelo recoge una serie de elementos que deben aparecer en nuestro instrumento de medida. ¿Con qué patrón oro podríamos establecer nuestro grado de validez?

Mientras escribo estas líneas se me ocurren tres (hay más): la opinión de un experto (obviamente "ciego" a las puntuaciones obtenidas); otra escala asentada en la bibliografía (se me ocurre la de Barthel de independencia en las actividades de la vida diaria); y por último con criterios que provienen directamente del modelo del que surge el instrumento y que son generalmente más difíciles de obtener (por ejemplo, en relación con una entrevista mucho más larga en la que se le pregunta a la persona y/o sus familiares por la presencia, cambio o ausencia de un conjunto de conductas).

El proceso sería en este caso que la persona en cuestión fuera valorado dos veces, bien en la misma ocasión o bien en ocasiones diferentes (Yo aquí lo haría en la misma ocasión: es más barato y en el caso de que fueran personas muy jóvenes o de edad avanzada se podrían confundir la validez con la evolución por el paso del tiempo). Con la puntuación obtenida en ambas valoraciones en una muestra amplia de sujetos (esta es otra: la muestra de sujetos en la que demuestro la validez de mi instrumento de medida debe ser amplia y representativa de la población en la que voy aplicar mi instrumento de medida) encuentro una correlación que en el caso de que sea satisfactoria sería un elemento de apoyo sobre la calidad de mi instrumento.

No quiero complicar la exposición mencionando otro tipo de validez, como la de contenido, pero debe quedar claro que un instrumento no valido es un instrumento inútil (además de una pérdida de tiempo para quien lo usa y para quien es medido por él).

Sobre la fiabilidad, indicar que con este nombre se designa hasta que punto la puntuación obtenida mediante la aplicación adecuada de un instrumento es precisa, o lo que es lo mismo, hasta que punto está libre de error. (No confundir con una aplicación inadecuada, por ejemplo, que no se cumpla el protocolo o que la persona dé información falsa). Esto significaría que si una persona utiliza el mismo instrumento de medida en dos ocasiones diferentes y esa dimensión no ha variado a lo largo del tiempo la puntuación debería ser la

misma. Esto no siempre se cumple y se asume que en toda ocasión se va a dar un error de medida (por ejemplo, por despistes del informante, o por inexactitudes en la respuesta).

Aunque hay diversos procedimientos para calcular este error de medida y que no vamos a entrar en ellos, con este error de medida llegamos a poder calcular a un intervalo en el que, asumiendo una distribución normal del mismo (daré por supuesto que todos sabemos lo que es una distribución normal. Es eso que disteis en Estadística y que tiene forma de curva invertida. Y si no os acordais, os garantizo que en cualquier libro de estadística aparece explicada, como por ejemplo en el clásico Amón(4)), se encontraría la puntuación verdadera de la persona.

Por ejemplo, supongamos que en nuestro instrumento sobre "autonomía personal" encontramos que el error de medida es 3. La puntuación de una persona en la misma es de 85: pues bien, el intervalo entre el cual se encontraría en el 95% de las ocasiones la puntuación real de esa persona sería la puntuación conseguida a la cual se le suma el producto del error de medida por 2 (el valor exacto sería $\pm$1,96, intervalo que incluye al 95% del área debajo de la curva normal. Echadle una visual a los apuntes de estadística). Aplicado a este ejemplo sería 85 $\pm$ (2 x 3), lo que nos daría un rango de puntuaciones entre 79 y 91. Lo que afirmaríamos es que con un 95% de posibilidades de acertar la puntuación real de esa persona se encontraría comprendida en tal intervalo.

De este modo un instrumento de medida válido, pero poco fiable (en ocasiones porque son pocos elementos y el cambio en las puntuaciones variando un único elemento es devastador, o porque la definición o redacción del elemento es confusa y muy sujeta a interpretaciones) pierde mucha de su utilidad a la hora de valorar una situación clínica y poder tomar medidas correctoras. Si el intervalo obtenido es amplio en demasía, poca luz nos puede aportar.

En resumen, por un lado tenemos la validez (hasta que punto este instrumento de medida mide lo que dice medir) y la fiabilidad (hasta que punto una medida es precisa comparada con el valor real en esa dimensión de esa persona en esa ocasión) como elementos necesarios para considerar la utilidad y la seriedad de un instrumento de medida. Con frecuencia leeremos descripciones de instrumentos sin información sobre la validez y la fiabilidad. Mi consejo es no hacer caso de tales instrumentos, ya que es difícil interpretar los resultados obtenidos. Y recordar algo ya mencionado: los instrumentos deben ser validados en la población que la que va a ser utilizado. Una traducción sin más de un instrumento validado en otro idioma simplemente no nos vale.

Sensibilidad y especificidad

Aquí voy a dar un pequeño salto. Supongamos que tengo un instrumento de medida ya demostrada su validez y fiabilidad. Lamentablemente aquí ya no puedo continuar con esa maravillosa e inventada escala de “autonomía

personal". (En el primero borrador sí continuaba con ella, pero resultaba confuso. Y encima el tiempo apremiaba... Fui a tiro seguro). Voy a poner como ejemplo el proceso de una escala real como fue la adaptación al castellano del EAT (5) hecha por Castro, Toro, Salamero y Guimerá en 1991. (No me cito a mí mismo: se trata de Josefina Castro, psicóloga clínica cuyos estudios son fundamentales dentro del campo de los trastornos de la alimentación en España. Y del resto de los firmantes no se puede decir menos).

Brevemente descrito, el EAT es un cuestionario de 40 items que es cubierto por la persona tras una breve explicación. Inicialmente construido por Garner y Garfinkel (1979) para evaluar actitudes y conductas encontradas en pacientes con anorexia.

En el trabajo mencionado el instrumento fue cubierto en la consulta inicial, donde se aplicaban igualmente los criterios DSM-III (6) para el diagnóstico de Anorexia Nerviosa. Este diagnóstico, realizado por clínicos, fue tomado como el criterio externo respecto al cual se va a chequear la puntuación del EAT.

Otro elemento, amén del criterio externo, es a partir de qué puntuación se va a considerar la puntuación de una persona como caso o no caso clínico. Esta puntuación se denomina punto de corte, y es un elemento crítico: dependiendo del mismo se obtendrán más o menos casos.

Entonces tenemos por un lado un criterio externo y por el otro el punto de corte. Las definiciones a comentar a continuación son las de sensibilidad y

especificidad. Pero primero tomemos el cuadro que se menciona en este trabajo (página 181). En el mismo se recogen los datos de un grupo de personas con anorexia nerviosa y un grupo control, equiparado en sexo, edad y clase social. El criterio externo, decía, era el diagnóstico DSM-III asignado por un clínico experto. El punto de corte era la consideración de caso para puntuaciones mayores de 30.

	Grupo		
	Anorexia	Control	Total
EAT > 30	53	11	64
EAT ≤ 30	25	67	92
Total	78	78	156
Sensibilidad	53/78 (67,9%)	Especificidad	67/78 (85,9%)
Valor Pro. Positivo	53/64 (82,8%)	Valor Pro. Negativ	67/92 (72,8%)

En esta tabla aparecen otros dos conceptos (los de valores pronóstico negativo y positivo, que después diré qué son). Vayamos por los primeros. La sensibilidad nos indica la proporción o porcentaje de personas enfermas (o con cualquier otra característica que nos interese detectar) que podemos localizar con nuestro instrumento. Si localizáramos todas las personas enfermas, sería una sensibilidad del 100% (o 1, si es una proporción). En este caso la sensibilidad con este punto de corte es del 67,9%. ¿Qué quiere decir esto? Pues que en teoría en una muestra cualesquiera de población con este

instrumento y este punto de corte podríamos detectar a casi el 68% de todos los casos clínicos. Aquí vemos que no sólo el instrumento importa: también el punto de corte es crítico. De hecho situando el punto de corte en 20 la sensibilidad subiría hasta el 91% (este cuadro aparece también en la misma página 181 del trabajo mencionado).

Pasemos a intentar explicar lo que es especificidad. En este caso lo que me interesa es ya no detectar los "enfermos": lo que me interesa es encontrar los "sanos" (por llamarlos de una manera). La especificidad me indica el porcentaje o proporción de sanos que son detectados por el instrumento. En este caso con este instrumento *y* este punto de corte la especificidad es del 85,9% (o de 0,859). Esto significa que de cada 100 sanos detectaría a 86. ¿Y como afecta aquí el punto de corte? Pues de igual manera que en el párrafo anterior un punto de corte de 20 afectó en la sensibilidad subiendo ésta del 68% al 91%, en el caso de la especificidad ésta baja de casi el 86% al 69,2%. Esto es una constante: cuando el punto de corte es menos exigente encontramos más "enfermos o casos", pero también detectamos menos "sanos o no casos".

Pero en ocasiones detectamos instrumentos que tienen una alta sensibilidad, pero una especificidad tan baja que afecta en realidad a todo el instrumento y lo convierte en poco útil. Por ejemplo, en un capitulillo (por su extensión) de Hoes (7) también del año 1991 (parece que estoy abonado a este año, pero ha sido una simple casualidad: estaba leyéndolo sin un objetivo concreto

cuando encontré este ejemplo ¡Las musas a veces nos dan estas oportunidades!) encuentro que un instrumento, que denomina HVS (en realidad una escala con una serie de criterios y un punto de corte pre-establecido sobre hiperventilación) podría ser útil para el diagnóstico del trastorno de angustia según criterios DSM-III. En la página 63 encontramos la siguiente tabla (que modifico para homologarla con la anterior. Tal y como venía era me resultaba poco clara).

	Criterios diagnósticos DSM-III		
	T. Angustia	No T. Angustia	Total
HVS positivo	59	111	170
HVS negativo	5	99	104
Total	64	210	274
Sensibilidad	59/64 (92,2%)	Especificidad	99/210 (47,1%)
Valor Pro. Positivo	59/111 (34,7%)	Valor Pro. Negat.	99/104 (95,2%)

Aquí podemos comprobar como este instrumento si bien tiene una alta sensibilidad (92,2%), su especificidad es bastante baja (47,1%). Esto le quita mucha de la utilidad que podría tener como instrumento diagnóstico per se.

Ahora voy a referirme a la línea inferior: los Valores pronóstico negativo y positivo. Esto se informa de la probabilidad (o del porcentaje) de que un valor de “enfermo o caso” corresponda realmente a un “enfermo o caso” en el caso del valor pronóstico positivo. Y respecto al valor pronóstico negativo se refiere

a la probabilidad de que un “sano o no caso” esté realmente “sano” o sea un “no caso”. Estas propiedades, como las anteriores, dependerán de los puntos de corte establecidos: de ese modo cuando el punto de corte es más alto (o estricto) el valor pronóstico positivo tiende a aumentar y el valor pronóstico negativo a disminuir.

Volvamos al ejemplo de Castro y compañeros (5). En la tabla inicial se comprueba que el valor pronóstico positivo encontrado es de casi el 82,8%. Esto significa que de cada 100 positivos o presuntos casos “enfermos”, 83 están realmente “enfermos” (en este caso sufrían realmente anorexia). Por otro lado el valor pronóstico negativo es del 72,8%, lo cual quiere decir que de cada 100 negativos o presuntos “sanos”, 73 están realmente “sanos” (lo que quiere decir que en este caso, realmente no sufrían anorexia). Podemos comprobar que ambos valores están relativamente equilibrados.

En el segundo ejemplo, el de Hoes (7), lo que nos encontramos por un lado es un valor pronóstico negativo muy alto (del 95,2%) (o sea, que de 100 negativos, 95 no tenían trastorno de angustia), pero sin embargo un valor pronóstico positivo de tan solo un 34,7% (es decir, que de 100 positivos según la prueba, sólo 35 eran personas que realmente tenían trastorno de angustia y las otras 65 no).

¿Cómo conciliar ambos aspectos? Un modo de hacerlo es mediante la proporción de aquellos correctamente clasificados. Con estos dos ejemplos he de decir que el porcentaje de los correctamente clasificados en el primer

ejemplo es del 76,9% y en el segundo de 57,7%, 19 puntos menos. Parecería que el primer instrumento sería preferible al segundo (no es el caso ya que detectan cuadros diferentes. Pero asumamos que fuera el mismo), pero sobre esto no hay normas claras, y todo dependerá de la finalidad que el investigador o el clínico mantenga.

De este modo, si nos planteamos que en nuestro trabajo vamos a hacer al menos dos pasos, es muy posible que nos interese en un primer paso perder el mínimo número de casos o "enfermos" posible, aunque sea a costa de hincharse de falsos positivos, ya que los vamos a eliminar en un segundo paso (por ejemplo Aragonés Benaiges (8) en un trabajo del año 2005, y como prueba de algo posterior al año 1991 al que, insisto, no estaba abonado).

En otros casos nuestro objetivo puede ser diferente. En todo caso esto se sale de los límites impuestos a este artículo.

Conclusiones

El uso de instrumentos de medida en buena medida es necesario para la detección de problemas y/o para comprobar la evolución de un paciente, más allá de valoraciones subjetivas y no explícitas. Pero no cualquier conjunto de elementos es un instrumento de medida: deben probar que son válidos y fiables. De no ser así estaríamos obteniendo un número, pero no sabríamos de qué ni hasta qué medida este resultado refleja la realidad de la persona. Por lo tanto, cuidado con los instrumentos y no adoptar indubitablemente la

falsa seguridad que nos puede dar un número cuando ese número no es seguro.

Por último tenemos los términos de sensibilidad y especificidad. Esos términos tienen un significado diferente del coloquial: aquí son propiedades relacionadas no sólo con un instrumento de medida, sino también con un punto de corte en tal instrumento, y que éste puede depender de los objetivos que tenga el clínico o el investigador. Debemos tener eso en cuenta a la hora de escoger ambas cosas: instrumento y punto de corte.

BIBLIOGRAFÍA

1) Cronbach Lee J (1972): *Fundamentos de la exploración psicológica*, 2ª edición, Madrid, Biblioteca Nueva.

2) Muñoz Rodríguez P E (2000): *Tests de cribado en la práctica clínica,* en Medición clínica en psiquiatría y psicología (editado por Bulbena Vilarrasa A, Berrios G E & Fernández de Larrinoa Palacios P), pp 35- 54, Barcelona, Editorial Masson.

3) Cone, J D (1987): *Consideraciones "psicométricas" en la evaluación conductual.* En Evaluación conductual 3ª edición (editado por Fernández-Ballesteros & Carboles J A I), pp 159-184, Madrid, Ediciones Pirámide. 3ª edición

4) Amón J (1982): *Estadística para psicólogos (Tomo I)*. 5ª edición, Madrid, Ediciones Pirámide.

5) Castro J, Toro J, Salamero M & Guimerá E (1991): *The eating attitudes test: validation of the spanish version*. Evaluación psicológica/psychological assessment 7(2), pp 175-190.

6) American Psychiatric Association (1985): *DSM-III Manual Diagnóstico y estadístico de los trastornos mentales.* 2ª reimpresión, Barcelona, Editorial Masson. (Titulo original "Diagnostic and statistical manual of mental disorders, third edition", 1980).

7) Hoes, M J A J M (1991): *Una revisión crítica de la farmacoterapia antidepresiva: promesas y realidades*. En Realidad actual y perspectivas en el tratamiento farmacológico de la depresión (Editado por Ledesma Jimeno A & Prieto Aguirre J F), pp 55-70. Barcelona, ESPAXS.

8) Aragonés Benaiges E (2005): Estudio de los trastornos depresivos en la atención primaria de salud. Revista de Psiquiatría de la Facultad de Medicina de Barcelona, 32(1), pp 30-37.